



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Computational Grammar Acquisition from CHILDES Data Using a Probabilistic Parsing Model

Citation for published version:

Kwiatkowski, T, Goldwater, S & Steedman, M 2009, Computational Grammar Acquisition from CHILDES Data Using a Probabilistic Parsing Model. in *Psychocomputational Models of Human Language Acquisition (PsychoCompLA 2009)*.

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Peer reviewed version

Published In:

Psychocomputational Models of Human Language Acquisition (PsychoCompLA 2009)

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



Computational Grammar Acquisition from CHILDES data using a Probabilistic Parsing Model.

Tom Kwiatkowski, Sharon Goldwater, Mark Steedman

School of Informatics, University of Edinburgh

t.m.kwiatkowski@sms.ed.ac.uk, {sgwater, steedman}@inf.ed.ac.uk

1 Introduction

In this work we propose a universal model of syntactic acquisition that assumes the learner is exposed to pairs consisting of strings of word-candidates and contextually-afforded meaning-representations.

Previous attempts to model the learning of syntax (Siskind 1992, 1995, 1996; Villavicencio 2002; Yang 2002; Buttery 2003) have tended to adopt a “parameter-setting” approach (Hyams 1986; Gibson & Wexler 1995; Fodor 1998). However, recent work in the related task of inducing a grammar from a corpus of paired English sentences and database queries (Zettlemoyer & Collins 2005, Zettlemoyer & Collins 2007, Wong & Mooney 2007, Lu et al. 2008) has shown that it is possible to learn grammars without this “switch like” mechanism by using the structure of the meaning representation to bootstrap the syntactic learning procedure.

The present paper shows that these related methods can be generalized to provide a universal model of child language acquisition and our model is designed to be psycholinguistically plausible: the initialisation of the grammar is language independent and should be able to learn any plausible word order; and the model learns in a sequential manner from sentence - meaning pairs. For the purposes of this paper, we present only the case of learning from unambiguous sentence-meaning pairs. However, the principles used will extend to the case of learning in the face of spurious distracting meaning candidates that are contextually supported but irrelevant to the utterance.

2 Logical Form

Sagae et al. 2007 have recently annotated a substantial part of the English section of the CHILDES database with dependency graphs of the kind illustrated in figure 1. While this annotation scheme was designed to represent syntactic relations, these dependency graphs can be viewed as impoverished logical forms representing pure predicate-argument meaning relations, provided that the following language-specific aspects of the annotation are ignored by the learner. First, the learner must make no use of the fact that the dependency graph aligns the terminals of the predicate argument structure with words of English in an English sentence. For example, the learner must consider the possibility that the unknown word “blocks” corresponds to the semantic predicate *get* in figure 2.

Second, the learner must also ignore the fact that the mapping from nodes in the dependency graph to En-

glish words is one-to-one. For example, it should consider the possibility that the word “get” corresponds to the compound meaning abbreviated as *get out*.

Third, the learner must map dependency graphs like figure 1 onto structured logical forms like figure 2, in which terms must first be distinguished as functors, arguments, or adjuncts, so that they can be semantically typed.

We can assume that POS tags like *NN*, *VP* and directional dependencies labeled with relations like *jct* in dependency graphs like (1) can be mapped by rule in this way onto semantic types which for mnemonic reasons and ease of reading we will represent as basic unlinearized category schemata *S*, *NP*, *S\NP*, etc.: These type-schemata should be thought of as primarily semantic in nature, distinct from directional syntactic types like *S*, *NP*, *S\NP*, etc. that instantiate them for a particular language. The full set of such type schemata is given in figure A-1.

3 CCG Universal Grammar

A Combinatory Categorical Grammar consists of a language-specific lexicon whose entries are triples $\langle \text{word} := \text{syntactic category} : \text{logical form} \rangle$, and a universal set of syntactic combinatory rules that project the lexicon of a language onto its sentences.

For example, the English lexicon includes the following entries:

the	:=	<i>NP/N</i>	:	<i>the</i>
blocks	:=	<i>N</i>	:	<i>blocks</i>

The syntactic type *NP/N* identifies English “the” as combining with nouns of type *N* to its right to yield NPs. The corresponding lexical entry in a determiner-final language such as Lakhota would be written $\text{ki} := \text{NP}\backslash\text{N} : \text{the}$. The logical form *the* is a place-holder for the presumed universal semantics of definites, which may or may not be separately lexicalized in any given language.

The present paper uses only the rules of Application and Harmonic Composition, illustrated in figure A-2 as a result of which, the present system can only learn languages that are weakly context-free. However, it will generalize to the trans-context-free set covered by full CCG.

Consider the case in which a child equipped with the above universal rules but with no lexicon at all hears the sentence MORE DOGGIES! and knows unambiguously that this means *more dogs*. She can ap-



Figure 1: Syntactic dependency graph

$$[S_q \text{ can}_{(S_q|NP_{subj})|(S_{inf}|NP)}] [S_{inf}|NP \text{ get}_{(((S_{inf}|NP)|NP)|(S_{nom}|NP))} \text{ out}_{S_{nom}|NP}] [NP \text{ the}_{NP|N} \text{ blocks}_N]] \text{ you}_{NP}]$$

Figure 2: pseudo logical form

ply the universal combinatory rules in reverse to the pair $NP : \text{more dogs}$ to directly generate all possible ways that universal rules could project all possible lexical entries, pairing all possible words with all possible decompositions of the logical forms. As the only two combinatory rules that have a non-function category as their result are the rules of function application, the type-and-meaning representation $NP : \text{more dogs}$ generates just three derivations, illustrated in A-1.3.

Of these, the first derivation is correct for determiner-first languages like English. The second would be correct for a determiner-final language like Lakota. The third would be correct for a language where *more dogs* was realized as a single word.

4 Model

We use a probabilistic parsing model to generate all candidate parses for each sentence/logical-form pair in the training set. This model, described in A-1.4, works by first generating a syntactic parse tree with CCG syntactic categories at the nodes before then generating associated components of logical-form and words at the leaves of this tree. The model makes use of the conjugate-exponential Dirichlet Distribution and Dirichlet Process priors and is trained using the online Variational Bayesian Expectation Maximisation algorithm (Beal (2003)). This training procedure is online in the strong sense that each training pair is seen sequentially and only once.

5 Experiments

The model is trained on a set of 3599 child-directed sentence; logical-form pairs from the first 15 files of the Eve corpus discussed in section 2. These were collected between the ages of 1;6 and 2;1 (years; months) and only those sentences of 6 words or fewer were used, giving 10^4 word candidates for which the universal grammar licenses 2×10^5 distinct $\langle \text{word}, \text{meaning}, \text{syntactic category} \rangle$ triples. Our test set is made up of the child-spoken sentence; logical-form pairs from files 14 and 15 of the Eve corpus (collected at age 2;1).

Our evaluation is similar to that used in the semantic parsing literature, where the parsing model is used to predict logical-forms for a test set of sentences. We score these predicted logical-forms against the gold standard logical-forms with which the test sentences are annotated, reporting both exact-match accuracy

and partial-match accuracy, where the latter relates to the directed, labelled, dependencies within the logical-forms.

Table 1 gives precision, recall and f-score for both exact-match accuracy and partial match accuracy. Results are reported for the full test set and also for the subset (79%) of the test set which contains only words that were observed in the training set. These results

Words seen in training set				
		Precision	Recall	f-score
exact-match	baseline	100	13.6	23.9
	model	62	36	45.5
partial-match	baseline	100	19	31.9
	model	70	74	71.9
Full training set				
		Precision	Recall	f-score
exact-match	baseline	100	10	18.2
	model	51	28	36.2
partial-match	baseline	100	16.3	28.0
	model	61.9	67.5	64.6

Table 1:

show the parsing model significantly outperforming the baseline of memorised seen sentence-meaning pairs indicating an accurate lexicon and grammar. It should be noted that the training data for our model constitutes only a small subset of the child's full linguistic exposure (34 hours over a 7 month period). We expect would perform with a much higher accuracy if it were given a training set of a comparable size to that available to the child.

6 Conclusion

The above account represents the first step in building a universal model of first language acquisition. We have shown that there is a general method for mapping strings of English paired with impoverished meaning representations derived from dependency annotations onto a grammar/parser that builds such knowledge representations, without any English-specific language engineering, and that the parser trained on a subset of the Eve corpus in a psycholinguistically plausible online manner has built a reasonably accurate model of the CCG lexicon and grammar on the basis of a very small amount of data.

A-1 Supporting Material

A-1.1 Type schemata

$S_{[decl]}$: for declarative sentences
 $S_{[wh]}$: for wh questions
 $S_{[q]}$: for Yes/No questions
 $S_{[to]}|NP_{SUBJ}$: for to-infinitives
 $S_{[b]}|NP_{SUBJ}$: for bare-infinitives
 NP_{SUBJ} : for subject noun phrases
 NP_{OBJ} : for object noun phrases
 NP_{PRED} : for predicate noun phrases
 NP : for noun phrases
 N : for nouns
 PP : for prepositional phrases

Figure A-1: Semantic type schemata

A-1.2 CCG combinators

Application

$X:f(a) \rightarrow X/Y:\lambda x.f(x) \ Y:a$
 $X:f(a) \rightarrow X\backslash Y:\lambda x.f(x) \ Y:a$

Harmonic Composition

$X/Z:\lambda x.f(g(x)) \rightarrow X/Y:\lambda x.f(x) \ Y/Z:\lambda x.g(x)$
 $X\backslash Z:\lambda x.f(g(x)) \rightarrow Y\backslash Z:\lambda x.g(x) \ X\backslash Y:\lambda x.f(x)$

Figure A-2: CCG combinators

A-1.3 Parse forest

The type-and-meaning representation $NP : more \ N : dogs$ generates just three derivations:

- a.
$$\frac{\frac{MORE}{NP/N : more'_{((e,t),e)}} \quad \frac{DOGGIES}{N : dogs'_{(e,t)}}}{NP : more' dogs'_e} >$$
- b.
$$\frac{\frac{MORE}{N : dogs'_{(e,t)}} \quad \frac{DOGGIES}{NP\backslash N : more'_{((e,t),e)}}}{NP : more' dogs'_e} <$$
- c.
$$\frac{MORE \ DOGGIES}{NP : more' dogs'_e} !$$

A-1.4 Parsing Model

In order to generate a parse, the model first generates a syntactic parse tree with CCG syntactic categories at the nodes before then generating associated components of logical-form and words at the leaves of this tree.

We denote a single syntactic node in the parse tree as σ , a single node representing a component of logical-form as λ and a single word node as ϕ .

The generative process used to generate a string of words and associated component of logical-form is illustrated in figure A-1.4 (for which we have borrowed elements of the notation of Liang et al. (2007) since - as they point out - there is no convenient way of representing parse trees in the visual language of traditional graphical models).

This process proceeds by first drawing the top node of the parse tree (σ_{top}) from a Multinomial distribution over the atomic syntactic categories. We then build the tree by recursively drawing either a pair of syntactic children ($\langle \sigma_{l(i)}, \sigma_{r(i)} \rangle$) or a lexical item from each syntactic node σ_i in the parse tree.

In order to decide whether to generate a pair of syntactic children or lexical item for each syntactic node σ_i , the model draws a binary *rule-type* variable (t_i) from a Binomial distribution. If this variable licenses a syntactic expansion then the syntactic children of σ_i are drawn from a Multinomial distribution covering all the possible expansions of σ_i according to the universal grammar.

Alternatively, if t_i indicates that σ_i is a leaf node in the parse then a component of logical form λ_i is drawn from a Multinomial conditioned on the category σ_i and a word ϕ_i is then drawn from a Multinomial conditioned on the $\langle \sigma_i, \lambda_i \rangle$ pair.

We define priors for each of the Multinomial distributions used in this procedure and in the generation of a single parse, the parameters of the Multinomials are drawn from these priors (note that the Binomial distribution is a special case of the Multinomial).

For the Multinomial distributions used in producing the top syntactic node of the tree; the syntactic children of each non-terminal syntactic node; and the *rule-type* variables, we assign conjugate-exponential Dirichlet priors.

For the Multinomial distributions used to generate the components of logical-form and the words however, we cannot use the Dirichlet Distribution prior as the full scope of the lexicon cannot be known to the child (and therefore to our model) before the start of the language acquisition procedure. For these distributions we then use the infinitely expandable (but still conjugate-exponential) Dirichlet Process as a prior.

It should be observed that none of the nodes in figure A-1.4 are observed as we do not know the correct segmentation of either the sentence or the sentential logical-form. However, given the syntactic derivation created and the linear order of the leaf nodes, there is a deterministic (probability 1) mapping between the elements $\phi_i, \lambda_i : i = 1 \dots N$ and the observed pair (S,I), we have just chosen not to depict this in figure A-1.4 for reasons of clarity.

